Measuring Metagenome Assembly Quality in IMG

IMG now provides two metrics to evaluate the assembly quality of metagenomes and metagenome-like objects (metatranscriptomes, single particle sorts and cell enrichments). Users will be able to view the assembly quality information from both the metagenome list display and the metagenome detail pages.

Assembly Quality Metrics based on read mapping (JGI-generated data only)

The first method depends on the standardized JGI assembly pipeline, and it is only provided for JGI sequenced metagenomes, with the exception of some of the oldest dataset (prior to ~ 2015).

JGI assembly pipeline reports and stores the number of aligned reads (or mapped reads) and the number of input reads and/or quality-filtered reads).

The reported assembly quality data in AMG include:

- Number of Filtered Reads: number of reads available after quality filtering
- Total Filtered Bases: number of bases in the filtered reads
- Number of Mapped Reads: number of reads aligned to the assembly
- Average Read Length: = (Total Filtered Bases) / (Number of Mapped Reads)
- Total Mapped Bases: = (Number of Mapped Reads) * (Average Read Length)
- Average Coverage of Assembled Sequences: = (Total Mapped Bases) / (Number of Bases in the assembly, AKA “Genome Size”)

Assembly Quality Metrics based on Single Copy Marker Pfams

The second method provides an estimation of the number of genomes in a metagenome assembly based on counts of single copy marker genes, and therefore it is applicable to all metagenomes in IMG. This approach is based on the “contig-stats” functionality from Anvi’o (described at http://merenlab.org/2015/12/07/predicting-number-of-genomes/). IMG uses a list of 139 marker genes identified in Campbell et al., 2013 (https://doi.org/10.1073/pnas.1303090110). Each marker gene is represented by a unique PFAM domain, hence IMG annotation are mined based on the following list:

```
marker_pfams = [
  "pfam03485", "pfam03484", "pfam01121", "pfam03772", "pfam03602", "pfam06418",
  "pfam02224", "pfam00712", "pfam02767", "pfam02768", "pfam00035", "pfam00889",
  "pfam01176", "pfam00113", "pfam03952", "pfam06574", "pfam03147", "pfam01687",
```

"pfam02938", "pfam02527", "pfam00958", "pfam01025", "pfam01018", "pfam11987",
"pfam04760", "pfam00707", "pfam05198", "pfam01715", "pfam06421", "pfam01795",
"pfam02873", "pfam08529", "pfam02410", "pfam01195", "pfam01252", "pfam00162",
"pfam02912", "pfam03726", "pfam01416", "pfam02033", "pfam00154", "pfam02132",
"pfam00825", "pfam00687", "pfam00466", "pfam00298", "pfam03946", "pfam00542",
"pfam00572", "pfam00238", "pfam00252", "pfam01196", "pfam00828", "pfam00861",
"pfam01245", "pfam00181", "pfam03947", "pfam00453", "pfam00829", "pfam00237",
"pfam00276", "pfam01016", "pfam00830", "pfam00831", "pfam00297", "pfam01783",
"pfam01632", "pfam00573", "pfam00281", "pfam00673", "pfam00347", "pfam03948",
"pfam01281", "pfam00338", "pfam00411", "pfam00164", "pfam00416", "pfam00312",
"pfam00886", "pfam00366", "pfam01084", "pfam00203", "pfam00318", "pfam01649",
"pfam00189", "pfam00163", "pfam00333", "pfam03719", "pfam01250", "pfam00177",
"pfam00410", "pfam00380", "pfam01782", "pfam01000", "pfam03118", "pfam01193",
"pfam04997", "pfam00623", "pfam04983", "pfam05000", "pfam04998", "pfam04563",
"pfam04561", "pfam04565", "pfam10385", "pfam00562", "pfam04560", "pfam01765",
"pfam07499", "pfam01330", "pfam05491", "pfam02773", "pfam02772", "pfam00584",
"pfam03840", "pfam00344", "pfam02403", "pfam01668", "pfam02978", "pfam00763",
"pfam02882", "pfam00121", "pfam08275", "pfam03461", "pfam05698", "pfam05697",
"pfam01746", "pfam00750", "pfam01409", "pfam01509", "pfam00627", "pfam02130",
"pfam02367", "pfam03652", "pfam12344", "pfam08459", "pfam10458", "pfam06071",
"pfam06689"]

Each metagenome annotated by IMG will have a total gene count associated with each Pfam ID (some counts can be 0). IMG then computes the Mean, Median, Mode and Standard Deviation of these gene counts across the 139 single copy marker Pfams.

The reported assembly quality data is computed as follows:

- Mean: mean value of the 139 gene counts
- Median: median value of the 139 gene counts
- Mode: mode of the 139 gene counts – It is possible that there is no Mode, depending on the underlying distribution of counts.
- Std Dev: standard deviation of the 139 gene counts
- Est. Num. of Genomes:

    - set to the value of the Mode, if Mode exists;
    - otherwise set to the Median.
- Est. Avg. Genome Size: = (Number of Bases in the assembly AKA “Genome Size”) / (Est. Num. of Genomes)

There is an additional checking on the Coefficient of Variation (= (Std Dev) / (Mean)). If the coefficient of variation is greater than or equal to 50%, then we consider that this approach does not provide a reliable estimation of the number of genomes in the assembly, and IMG does not report estimated number of genomes or estimated average genome size.

Finally, we expect that the estimated average genome size should be between 1 Mb and 10 Mb for most standard metagenomes, and IMG reports “Unusually low/high genome size estimated, this dataset may not be a standard microbial metagenome.” if the estimated average genome size is not within this expected range.